Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 3565–3571
Bioorganic & Medicinal Chemistry

2D-autocorrelation descriptors for predicting cytotoxicity of naphthoquinone ester derivatives against oral human epidermoid carcinoma

Liane Saíz-Urra,[a] Maykel Pérez González[a,b,c,*] and Marta Teijeira[b]

[a]*Chemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, C.P. 54830, Cuba*
[b]*Department of Organic Chemistry, Vigo University, C.P. 36200, Vigo, Spain*
[c]*Service Unit, Experimental Sugar Cane Station ''Villa Clara-Cienfuegos'', Ranchuelo, Villa Clara, C.P. 53100, Cuba*

Received 15 January 2007; revised 16 February 2007; accepted 19 February 2007
Available online 22 February 2007

Abstract—A QSAR study was developed, employing 2D-autocorrelation descriptors and a set of 37 naphthoquinone ester derivatives, in order to model the cytotoxicity of these compounds against oral human epidermoid carcinoma (KB). A comparison with other approaches such as the BCUT, Galvez topological charge indexes, Randić molecular profile, Geometrical, and RDF descriptors was carried out. Mathematical models were obtained by means of the multiple regression analysis (MRA) and the variables were selected using genetic algorithm. Based on the statistical results the 2D-autocorrelation descriptors were considered the best and were able to describe more than 84.2% of the variance in the experimental activity once we controlled for outliers.

1. Introduction

The search for new drugs against cancer plays a central role in the research programs of pharmaceutical companies and many governmental organizations due to the impact of this disease.

Anticancer activity of several types of naphthoquinones has been reported previously, for example against Walker 256 Carcinosarcoma in rats.[1,2] The naphthoquinone derivatives also display antiproliferative activity on HeLa cells as expected for CDC25 inhibitors and inhibit cell growth in a clonogenic assay at submicromolar concentrations. They increase inhibitory tyrosine 15 phosphorylation of CDK and induce the cleavage of PARP, a hallmark of apoptosis.[3] On the other hand, the 3-chlorodeoxylapachol is a naphthoquinone from *Avicennia germinans* which is cytotoxic in a panel of human cancer cells, and active against oral human epidermoid carcinoma (KB) in the murine hollow fiber antitumor model, with selectivity in KB cells for the intravenous site at lower doses, indicating possible metabolic activation.[4]

Another interesting group of compounds contains Rhinacanthins, which are naphthoquinone ester derivatives isolated from methanolic extract of the roots of the medicinal plant *Rhinacanthus nasutus* and have been reported because of their cytotoxicity against P388, A549, HT-29, and HL-60 cell lines.[5]

The aim of this study is to develop a QSAR model of the cytotoxicity of naphthoquinone ester derivatives against KB, to better understand the structural features of these types of compounds and their relation with the anticancer activity using for the first time the 2D-autocorrelation descriptors. This study may help us to design new analogues with better biological profile.

2. Materials and methods

2.1. Data set

Our data set consisted of 37 naphthoquinone esters that can be seen in the supplementary material with their

Keywords: QSAR; Naphthoquinone esters; Anticancer activity; 2D-autocorrelation descriptors; Cluster analysis.
* Corresponding author. Tel.: +53 42281473; e-mail: mpgonzalez76@yahoo.es

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.02.032

50% inhibition concentration (IC_{50}) observed in KB cell lines.[5] We used 37 compounds of the total set of 42 reported by Kongkathip et al.,[5] because five naphthoquinones were excluded due to an inexact IC_{50} concentration and therefore could not be used in the multiple regression analysis. To guarantee the linear distribution of the dependent variable(s), we calculated the natural logarithm of the IC_{50} values and constructed the model using these values.

2.2. Computational strategies

For these compounds we carried out geometry optimization calculations using the quantum chemical semi-empirical method AM1[6] implemented in MOPAC 6.0.[7] The DRAGON[8] computer software was employed to calculate all of the molecular descriptors in this study.

The mathematical models were obtained by means of Multiple Regression Analysis (MRA) as implemented in the STATISTICA software.[9] The Genetic Algorithm (GA) was used in the variable selection strategy, in order to include in the equation the most significant parameters from the data set. The particular GA simulation conditions applied here are: 1000 generations, trade-off between crossover and mutation parameter was 0.5, 1 smoothness factor, and 300 model populations. The statistical significance of the models was determined by examining the squared regression coefficient (R^2), the standard deviation (S), and the Fisher ratio (F). The proposed models were also checked for reliability and robustness by permutation testing (Y scrambling). Evidence that the proposed model was not a result of chance correlation was provided by obtaining new models with significantly lower R^2 and q^2 than the original model. The orthogonalization process of molecular descriptors[10–14] was carried out to eliminate the collinearity among these ones. The Randić method of orthogonalization has been described in detail in several publications,[10–14] thus, we will only give a general overview here. The first step in orthogonalizing the molecular descriptors in a model is to select the appropriate order of orthogonalization, which, in this case, is the order in which the variables were selected in the genetic algorithm search procedure of the linear regression analysis. The first variable MATS6m is taken as the first orthogonal descriptor Ω^1MATS6m, and the second one is orthogonalized with respect to it by taking the residual of its correlation with Ω^1MATS6m. The process is repeated until all the variables are completely orthogonalized, and the orthogonal variables are then used to obtain the new model.

2.3. Validation of the models

The robustness of the models and their predictivity were evaluated by both q^2 'leave-one-out' (LOO) cross-validation and bootstrap tests. Nevertheless, in order to obtain validated QSAR models the data set was divided into the training and test sets (20% of the whole data). Ideally, this division is performed such that points representing both the training and test set are distributed within the whole descriptor space occupied by the entire data set, and each point of the test set is close to at least one point of the training set. K-means cluster analysis (k-MCA) was used in splitting the set of compounds to guarantee this distribution.[15–19]

2.4. Comparison with other approaches

On the other hand, the 2D-autocorrelation descriptors were compared with other methodologies such as BCUT,[20–22] Galvez topological charge indices,[23–26] Randic molecular profiles,[27,28] Geometrical,[29] and RDF descriptors.[30] The models relative to these methodologies were developed using the same data set. The comparisons were done based on the results of the regression analysis, the predictive capability of the models generated, and using other statistical criteria such as Akaike's information criterion (AIC; Eq. 1; the model that produces the minimum AIC value is considered potentially the most useful)[31,32] and Kubinyi function (FIT; Eq. 2; the best model will present the highest value of this function)[33,34]

$$\mathrm{AIC} = \mathrm{RSS} \cdot \frac{(n+p')}{(n-p')^2}, \tag{1}$$

$$\mathrm{FIT} = \frac{R^2 \cdot (n-k-1)}{(n+k^2) \cdot (1-R^2)}. \tag{2}$$

3. Results and discussion

The best QSAR model obtained with the 2D-autocorrelation descriptors is given below together with the regression results:

$$\begin{aligned} -\log(\mathrm{IC}_{50}) = {} & -0.934(\pm 0.162) \cdot MATS6m \\ & - 1.102(\pm 0.171) \cdot ATS1e \\ & + 13.632(\pm 2.109) \cdot GATS3p \\ & + 1.185(\pm 0.149) \cdot ATS8m \\ & - 13.590(\pm 2.048) \cdot GATS3v \\ & - 0.920(\pm 0.006), \end{aligned} \tag{3}$$

$$N = 30, \quad R^2 = 0.786, \quad S = 0.323, \quad F = 15.570,$$
$$p < 10^{-5}, \quad \mathrm{AIC} = 0.156, \quad \mathrm{FIT} = 1.590,$$
$$q^2_{\text{CV-LOO}} = 0.735, \quad q^2_{\text{boot}} = 0.701, \quad a(R^2) = 0.111,$$
$$a(q^2) = -0.265, \quad \mathrm{LOF} = 0.187,$$

where N is the number of compounds included in the model, R^2 is the square of the correlation coefficient, S is the standard deviation of the regression, F is the Fisher ratio, and p is the significance of the model. AIC is the Akaike's information criterion and FIT is the Kubinyi function. Furthermore, we calculated the validation parameters shown previously, including the cross-validated squared regression coefficient q^2 of the LOO ($q^2_{\text{CV-LOO}}$), bootstrapping (q^2_{boot}), and Y scrambling

($a(R^2)$ y $a(q^2)$) procedures. We also calculated Friedman's lack of fit (LOF) factor[35] which takes into account the number of terms used in the equation and is not biased, as are other indicators, toward large numbers of parameters. The equation relative to this parameter is given below

$$\mathrm{LOF} = \frac{\mathrm{RSS}/n}{\left[1 - \frac{k\cdot(d+1)}{n}\right]^2}, \tag{4}$$

where RSS is the residual sum of squares, n is the number of cases, k is the number of variables in the model, and d is smoothing parameter.

Although we determined that the 2D-autocorrelation descriptors family was statistically significant, we carried out a comparison of different methodologies to validate our model. The results obtained from this comparison are given in Tables 1 and 2. The meaning of the variables is presented in the supplementary material.

As can be seen in Table 1, the value of R^2 is lower than 0.653 for all methodologies except the 2D-autocorrelation which has an R^2 equal to 0.786. This methodology also yielded the best values for other statistical parameters like the standard deviation of the regression and the Akaike's information criterion which has the lowest values in comparison with the rest of the methodologies. In the same way, the Fisher ratio and the Kubinyi function are the highest.

Finally, all models require validation in order to demonstrate their predictive value; otherwise a QSAR is of no practical use. Statistical fit should not be confused with the ability of a model to make predictions. Taking into account the validation parameters, all methodologies had statistical results inferior to the results yielded by the 2D-autocorrelation. Note that the results of the cross-validated squared regression coefficient for the leave-one-out procedure yielded values lower than 0.50, the minimum value considered suitable,[36] except for the RDF descriptors. However, the value of the cross-validated squared regression coefficient for bootstrapping procedure is lower than 0.50. Therefore, the methodologies used in the comparison with the 2D-autocorrelation do not present a good predictive capability.

Additionally, the 2D-autocorrelation descriptors present the best Q^2_{EXT} (0.731) predicting the external test set regarding the better predictive power of the methodologies used. For all these reasons, we considered that these descriptors can be useful tools for the prediction of biological activity taking into account the naphthoquinone compounds.

Collinearity among variables was avoided using Randić's orthogonalization method minimizing the interrelatedness among them. The QSAR model obtained with the 2D-autocorrelation (Eq. 5) after orthogonalization and standardization is given below, together with the results of the regression analysis.

$$\begin{aligned} -\log(\mathrm{IC}_{50}) = {} & -0.126(\pm 0.066)\cdot {}^{1}\Omega MATS6m \\ & + 0.195(\pm 0.058)\cdot {}^{2}\Omega ATS1e \\ & - 0.176(\pm 0.066)\cdot {}^{3}\Omega GATS3p \\ & + 0.465(\pm 0.065)\cdot {}^{4}\Omega ATS8m \\ & - 0.372(\pm 0.057)\cdot {}^{5}\Omega GATS3v \\ & - 0.881(\pm 0.059), \end{aligned} \tag{5}$$

$$\begin{aligned} & N = 30, \quad R^2 = 0.786, \quad S = 0.323, \quad F = 15.570, \\ & p < 10^{-5}, \mathrm{AIC} = 0.156, \quad \mathrm{FIT} = 1.590, \\ & q^2_{\text{CV-LOO}} = 0.735, \quad q^2_{\text{boot}} = 0.701, \quad a(R^2) = 0.111, \\ & a(q^2) = -0.265, \quad \mathrm{LOF} = 0.187. \end{aligned}$$

To further test the QSAR model, it was important to examine the data outliers. The level of outliers can

Table 1. The statistical parameters of the linear regression models obtained for the six kinds of descriptors involved in the comparison

Descriptors	Variables	R^2	S	F	p	AIC	FIT
2D-autocorrelation	ATS8m, ATS1e, MATS6m, GATS3v, GATS3p	0.786	0.323	15.570	$<10^{-5}$	0.156	1.590
BCUT	BEHm7, BELv5, BELv6, BELp6, BELp8	0.447	0.518	3.880	$<10^{-4}$	0.403	0.352
Galvez topological charge indices	GGI1, GGI6, JGI2, JGI3, JGI4	0.409	0.535	3.330	$<10^{-3}$	0.430	0.375
Randić molecular profiles	DP07, DP14, DP15, SP17, SP18	0.558	0.463	6.050	$<10^{-4}$	0.322	0.461
Geometrical	J3D, H3D, MAXDP, RGyr, MEcc	0.535	0.475	5.51	$<10^{-5}$	0.339	0.447
RDF	RDF045m, RDF010v, RDF025e, RDF150e, RDF110p	0.652	0.411	8.890	$<10^{-5}$	0.254	0.588

Table 2. The validation parameters of the linear regression models obtained for the six kinds of descriptors involved in the comparison

Descriptors	Variables	$q^2_{\text{CV-LOO}}$	q^2_{boot}	$a(R^2)$	$a(q^2)$	LOF	Q^2_{EXT}
2D-autocorrelation	ATS8m, ATS1e, MATS6m, GATS3v, GATS3p	0.735	0.701	0.111	−0.265	0.187	0.731
BCUT	BEHm7, BELv5, BELv6, BELp6, BELp8	0.221	0.109	0.167	−0.305	0.483	0.103
Galvez topological charge indices	GGI1, GGI6, JGI2, JGI3, JGI4	0.193	0.081	0.144	−0.346	0.516	0.279
Randic molecular profiles	DP07, DP14, DP15, SP17, SP18	0.389	0.266	0.159	−0.363	0.386	0.163
Geometrical	J3D, H3D, MAXDP, RGyr, MEcc	0.284	0.146	0.125	−0.392	0.406	0.264
RDF	RDF045m, RDF010v, RDF025e, RDF150e, RDF110p	0.473	0.431	0.126	−0.363	0.304	0.301

become a serious problem because the model is unable to predict 'real' biological activity. In this context, we looked for the presence of outliers in Eq. 5. We extracted only one outlier which represented 3.33% of the data. The extraction of 10% of the general data is classically accepted in the literature as the threshold for this procedure. The two tests that we used to detect the presence of outliers were a standard residual higher than 2δ, where δ is equivalent to the standard deviation, and deleted residual.

The model obtained is given below with the structure of this outlier in Figure 1 and the new statistical parameters for the extraction in Table 3

$$\begin{aligned} -\log(IC_{50}) = & -0.126(\pm 0.067)\cdot {}^{1}\Omega MATS6m \\ & + 0.197(\pm 0.061)\cdot {}^{2}\Omega ATS1e \\ & + 0.177(\pm 0.059)\cdot {}^{3}\Omega GATS3p \\ & + 0.464(\pm 0.062)\cdot {}^{4}\Omega ATS8m \\ & - 0.372(\pm 0.058)\cdot {}^{5}\Omega GATS3v \\ & - 0.880(\pm 0.061). \end{aligned} \quad (6)$$

We made a comparison between compounds **2** and **11** in order to understand the outlier behavior of this compound, taking into account that compound **2** presents the highest activity value and compound **11** has the highest potential as outlier in data set. As can be seen in Figure 2 the only difference between these compounds is the substitution of a methyl group at the C-2 position in the propyl chain of the naphthoquinone ring by a hydrogen atom in compound **11** relative to compound **2**. In spite of this apparently small difference, compound **2** was 107-fold more active than compound **11**, displaying values of observed concentration IC_{50} (μM) equal to 0.35 ± 0.16 and 37.56 ± 1.34, respectively. However, the above-mentioned biological activities differed from the results predicted by our model which were 0.524 and 5.754, respectively, resulting in compound **2** being only 10.98-fold more active. In short, our model did not predict the high difference between the activities of these compounds resulting from their slight structural difference. This fact might result because the variables in our model were weighted by atomic mass and to a lesser extent by polarizability and atomic van der Waals volume. As we can see, the substitution of a methyl group by hydrogen atom in a molecule of this size is not a meaningful change in the calculation of the descriptors involved in the model taking into account these weights.

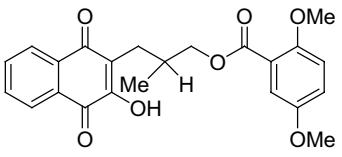

Figure 1. Chemical structure of the outlier found in the data.

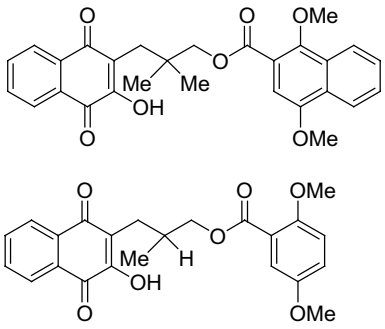

Figure 2. Comparison between compound **2** (upper panel) and compound **11** (lower panel).

On the other hand, compound **11** is a chiral compound and the observed value of concentration IC_{50} was reported taking into account the racemic modification, in other words, the reported IC_{50} is an average of the values of biological activity from both enantiomers. We think that the activity of each enantiomer could be quite different, being one of these (*R* or *S*) more active than the other due to a specific requirement of the configuration of the compounds to interact with the receptor.

In this connection, models were developed using 3D descriptors like RDF, in comparison with the 2D-autocorrelation descriptors, which identified compound **11** as a potential outlier too. This fact caused us to think that it would be very interesting to conduct a biological study about the anticancer activity of each particular enantiomer in order to better train the QSAR models. Finally, a table with the observed and predicted activity reported according to the model 6 relative to the 2D-autocorrelation descriptors and a figure showing the good correlation between predicted and observed anticancer activities according to this one are presented in the supplementary material.

Examining the molecular descriptors of the model we could get a better understanding of the relation between the structure of the compounds and their anticancer property. We focused on the final model which is described by Eq. 6.

The 2D-autocorrelation descriptors in general explain how the values of certain functions, at intervals equal to the *lag*, are correlated. In the case of the descriptors used to develop the model, *lag* is the topological distance *d* and the atomic properties are the functions correlated. There are slight differences between the 2D-autocorrelation descriptors ATS*d*, MAT*d*, GATS*d*; in general, they describe how the considered property is distributed along the topological structure. The most important

Table 3. Statistical parameters obtained when the outlier was removed from Eq. 5 for generating Eq. 6

Compound	R^2	S	F	AIC	FIT	q^2_{LOO}	Q^2_{boot}	$a(R^2)$	$a(q^2)$	LOF	Q^2_{EXT}
11	0.843	0.276	24.590	0.117	2.263	0.771	0.740	0.093	−0.372	0.141	0.743

factor in interpreting them in the model is the topological distance once weighted equally.[29]

As we can see in Eq. 6, the five most important descriptors among all the 2D-autocorrelation ones are weighted by the atomic mass, polarizability, electronegativity, and van der Waals volumes. The individual and progressive contributions of these descriptors to the model are shown next.

The 2D-autocorrelation descriptor most correlated with the anticancer activity was weighted by atomic masses which resulted in a 46.40% (R^2). Also the $^4\Omega ATS8m$ was the most important with a 33.0% (R^2). For that reason we are going to focus on this type of descriptors for our interpretation of the biological property.

The variable $^4\Omega ATS8m$ takes into account the interaction between each pair of atoms at the topological distance equal to 8 and weighted by the atomic mass. From the figures below it can be seen, broadly speaking, some of the possible interactions at *lag* equal to 8. These interactions involve, on the one hand, the chain of atoms between both ring systems along the entire molecule (see Fig. 3) and on the other hand the atoms in a single ring system. Possible fragments involved in this type of inter-atomic interaction are highlighted with a blue line in Figure 3.

These interactions are very important in the interpretation of the contribution of the variables in the model and the following figure will illustrate this graphically through a comparison between compounds **2** and **5**. Compound **2** is the most active as we demonstrated earlier and in Figure 3 it can be seen that there are many interactions at a lag equal to 8 in its structure. But in compound **5**, which is very similar to **2** in structure because it has the same substituents, we can see how instead of the naphthalene substructure found in compound **2**, there is only a benzene ring and fewer interactions at a lag equal to 8. The result is a lower value of IC_{50} (3.38 μM) in compound **5** with respect to compound **2** (IC_{50} = 0.35).

In the same way, this descriptor takes into account interactions that involve the methyl groups in the alkyl chain between both rings. The position of these methyl substituents allows them to interact with many atoms at *lag* equal to 8 in both rings increasing the value of this descriptor and then, the biological activity. This effect can be seen in the comparison between compounds **2** and **23** which only differ in the absence of the methyl substituents in the alkyl chain and the problem of predicting the activity for enantiomers is eliminated due to the absence of a chiral center (see the structure in supplementary material). Compound **2** results in being 37-fold more active than compound **23**, displaying values of observed concentration IC_{50} (μM) equal to 0.35 ± 0.16 and 12.78 ± 1.00, respectively (Fig. 4).

On the other hand, an important contribution to consider is with regard to the variable $^3\Omega GATS3p$ due to its positive coefficient in the Eq. 6 and its mathematical definition that implies only positive values, being its contribution always positive. A very interesting relation of this descriptor with the biological activity can be shown by the comparison between compounds **2** and **26**, since the main structural difference is the absence of the OH group in the naphthoquinone system of compound **26**. Compound **2** results in being 36-fold more active than compound **26**, displaying values of observed concentration IC_{50} (μM) equal to 0.35 ± 0.16 and 12.5 ± 0.98. The OH group linked to the naphthoquinone is at a topological distance equal to 3 from the oxygen atom of the carbonyl group. These atoms are the most polarizable and contribute significantly to increase the value of the $^3\Omega GATS3p$ and then, the biological activity.

As the training set only contains compounds with the OH linked to the naphthoquinone system at a topological distance equal to three to the oxygen of the carbonyl

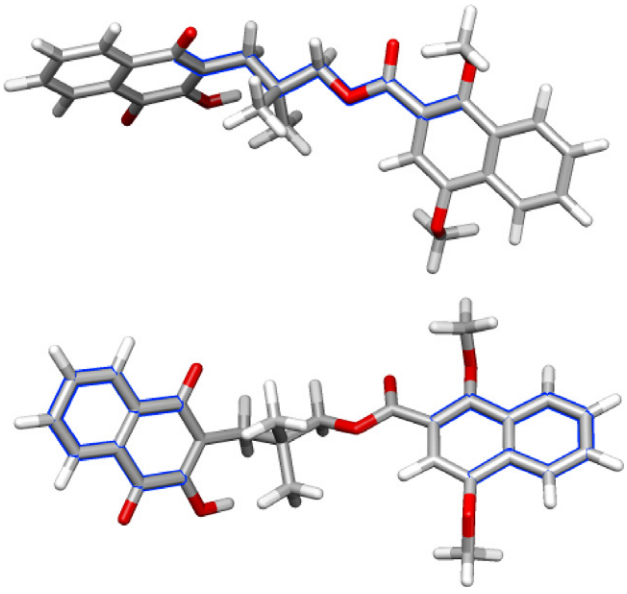

Figure 3. Fragments in blue represent possible inter-atomic interactions at *lag* equal to 8 in the compound **2**.

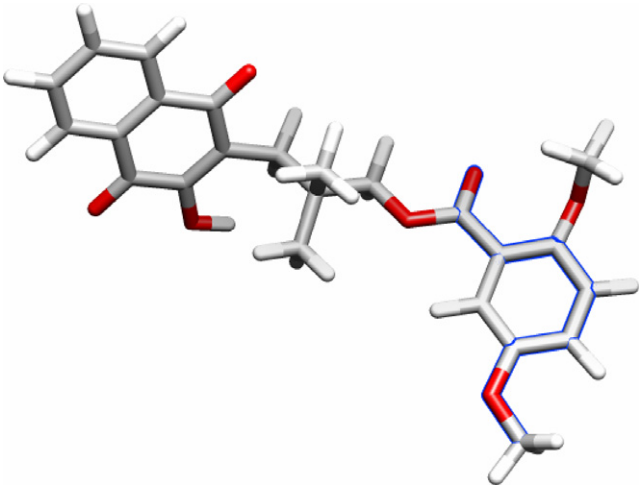

Figure 4. Fragment in blue represents a possible inter-atomic interaction at *lag* equal to 8 in the compound **5**.

group, it would be very interesting to predict the activity of designed compounds with the OH substituent in other positions in the same ring system by the model. With this aim, we kept the same structure of the most active compound (**2**) and moved the OH substituent to the other two possible positions that we found in the ring. The compounds are shown in Figure 5.

These compounds resulted in being 116.4- and 27.29-fold less active than compound **2** from the training set, displaying values of predicted concentration IC_{50} (μM) equal to 40.74 and 9.55, respectively.

Because of the information provided by the analysis above, we think that it would be interesting to design

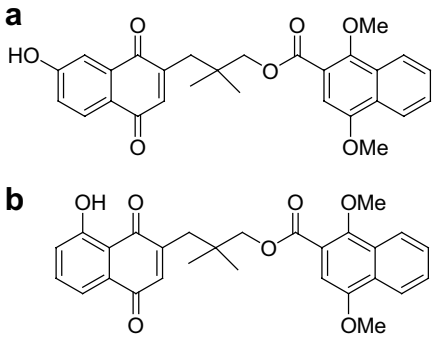

Figure 5. Designed compounds. (a) Compound **1d** and (b) compound **2d**.

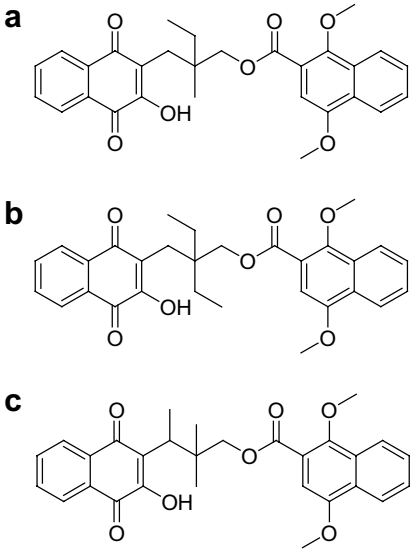

Figure 6. Designed compounds. (a) Compound **3d**, (b) compound **4d** and (c) compound **5d**.

a new naphthoquinone compound keeping the main core of compound **2**, with the naphthalene ring linked to the ester function and the OH substituent linked to the naphthoquinone system at a lag equal to 3 of the oxygen from the carbonyl group but changing the substituents in the alkyl chain. Three compounds (named **3d**, **4d**, and **5d**) were designed and their biological activity predicted according to the model. In the first one, a methyl substituent was replaced by an ethyl group; the second one has two ethyl substituents instead of two methyl ones, and finally the third one has one methyl linked to the alkyl chain between the naphthoquinone system and the two original methyl substituents. They can be seen in Figure 6.

The predicted IC_{50} of these designed compounds was 0.051, 0.058, and 0.209 μM displaying a potential activity 6.86-, 6.03-, and 1.675-fold higher than the pattern compound **2** (from the training set). The structural difference between compounds **3d** and **4d** is not significant taking into account that the position of the ramification in the alkyl chain is the same and the activity gap among these ones is very slight.

However, the addition of the methyl substituent in the position described above for the compound **5d** has other implications with regard to the biological activity. In spite of the lower value of the descriptor $^{4}\Omega ATS8m$ for this compound, the remarkable difference is the increase of the value of the descriptor $^{5}\Omega GATS3v$, being the second most important descriptor in the equation (see Table 4). Due to its mathematic definition and the coefficient in Eq. 6, its contribution is always negative; therefore, an increase of the ramifications in the alkyl chain is unfavorable for activity. This might be due to a possible steric impediment to rotate the alkyl chain to take certain conformation necessary for the interaction with the receptor.

4. Conclusion

In summary, in this study we modeled the anticancer activity, specifically the cytotoxicity of 30 compounds (IC_{50}) against oral human epidermoid carcinoma (KB). For this purpose we employed the 2D-autocorrelation descriptors.

The results produced by the methodology that we proposed were superior to the other descriptors such as BCUT, Galvez topological charge indexes, Randić molecular profile, Geometrical, and RDF; taking into account the statistical parameters of the model and the validation results.

Table 4. Statistical parameters obtained by the consecutive introduction of the variables in Eq. 6

Descriptors	R^2 progressive	R^2 individual	$q^2_{CV\text{-}LOO}$ progressive	$q^2_{CV\text{-}LOO}$ individual
$^{4}\Omega ATS8m$	0.330	0.330	0.253	0.253
$^{5}\Omega GATS3v$	0.538	0.208	0.466	0.213
$^{1}\Omega MATS6m$	0.662	0.124	0.594	0.128
$^{2}\Omega ATS1e$	0.761	0.099	0.685	0.091
$^{3}\Omega GATS3p$	0.843	0.082	0.771	0.086

The analysis of potential outliers suggested that the activity for the enantiomers could be different; where one of these (*R* or *S*) is more active than the other due to a specific requirement about the configuration of the compounds to interact with the receptor. Our model was not able to explain this point; in these cases the prediction of racemic modifications is similar to compounds without chiral centers and structural resemblance.

The variables that were found to be most significant in describing the model were weighted by atomic masses, polarizabilities, electronegativities, and atomic van der Waals volumes. The atomic mass is the most important weight and the Broto–Moreau autocorrelation of a topological structure distance equal to 8 was the most significant variable. This is illustrated in the higher activity of the compounds with a naphthalene ring in lieu of a benzene ring linked to the alkyl chain, and methyl substituents in the middle of this one. An increase in the ramification in the chain between both ring systems is unfavorable and it might be given by a possible steric impediment to rotate the alkyl chain to take certain conformation necessary for the interaction with the receptor. The second more important descriptor in the model, Geary autocorrelation at topological distance equal to 3 and weighted by atomic van der Waals volumes, describes this relation.

Finally, this is a simple model that might be used in the prediction of the anticancer activity against KB, of compounds related to the congeneric set of naphthoquinones as alternative and useful tool in the search of new anticancer compounds with better biological profile.

Acknowledgments

We acknowledge the Universidad de Vigo, Programa de Cooperación Inter-Institucional VLIR-UCLV and the Cuban Higher Education Ministry (R&D project number 6.181-2006) for financial support. Marta Teijeira thanks the Xunta de Galicia for the Parga Pondal contract. Last but not least the authors acknowledge Michel González for his useful comments for the development of this work.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2007.02.032.

References and notes

1. Tandon, V. K.; Yadav, D. B.; Chaturvedi, A. K.; Shukla, P. K. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3288.
2. Tandon, V. K.; Chhor, R. B.; Singh, R. V.; Rai, S.; Yadav, D. B. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1079.
3. Brun, M. P.; Braud, E.; Angotti, D.; Mondesert, O.; Quaranta, M.; Montes, M.; Miteva, M.; Gresh, N.; Ducommun, B.; Garbay, C. *Bioorg. Med. Chem.* **2005**, *13*, 4871.
4. Jones, W. P.; Lobo-Echeverri, T.; Mi, Q.; Chai, H.; Lee, D.; Soejarto, D. D.; Cordell, G. A.; Pezzuto, J. M.; Swanson, S. M.; Kinghorn, A. D. *J. Pharm. Pharmacol.* **2005**, *57*, 1101.
5. Kongkathip, N.; Luangkamin, S.; Kongkathip, B.; Sangma, C.; Grigg, R.; Kongsaeree, P.; Prabpai, S.; Pradidphol, N.; Piyaviriyagul, S.; Siripong, P. *J. Med. Chem* **2004**, *47*, 4427.
6. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
7. Frank, J. *Seiler Research Laboratory*; US Air Force Academy: Colorado Springs CO, 1993.
8. Todeschini, R.; Consonni, V.; Pavan, M., Dragon Software version 2.1, 2002.
9. Statsoft, Inc. STATISTICA (data analysis software system) version 6.0, 2002. Available at www.statsoft.com.
10. Klein, D. J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. *Int. J. Quant. Chem.* **1991**, *63*, 215.
11. Randić, M. *J. Mol. Struct. (Theochem.)* **1991**, *233*, 45.
12. Randić, M. *New J. Chem.* **1991**, *15*, 517.
13. Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
14. Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532.
15. González, M. P.; Helguera, A. M.; Cabrera, M. A. *Bioorg. Med. Chem.* **2005**, *13*, 1775.
16. González, M. P.; Dias, L. C.; Helguera, A. M.; Rodriguez, Y. M.; de Oliveira, L. G.; Gomez, L. T.; Diaz, H. G. *Bioorg. Med. Chem.* **2004**, *12*, 4467.
17. Molina, E.; Diaz, H. G.; González, M. P.; Rodriguez, E.; Uriarte, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 515.
18. González, M. P.; Gonzalez Diaz, H.; Molina Ruiz, R.; Cabrera, M. A.; Ramos de Armas, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192.
19. González, M. P.; Terán, C.; Fall, Y.; Diaz, L. C.; Morales, A. H. *Polymer* **2005**, *46*, 2783.
20. Burden, F. R. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225.
21. Burden, F. R. *Quant. Struct.- Act. Relat.* **1997**, *16*, 309.
22. Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28.
23. Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. *J. Chem. Inf. Comp. Sci.* **1994**, *34*, 520.
24. Galvez, J.; Garcia-Domenech, R.; de Gregorio Alapont, C.; de Julian-Ortiz, J. V.; Popa, L. *J. Mol. Graph.* **1996**, *14*, 272.
25. Galvez, J.; Garcia-Domenech, R.; de Julian-Ortiz, J. V.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272.
26. Rios-Santamarina, I.; Garcia-Domenech, R.; Galvez, J.; Cortijo, J.; Santamaria, P.; Morcillo, E. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 477.
27. Randic, M. *New J. Chem.* **1995**, *19*, 781.
28. Randic, M. *J. Chem. Inf. Comp. Sci.* **1995**, *35*, 373.
29. Todeschini, R.; Consonni, V. In *Handbook of Molecular Descriptors*; 1. Edition ed.; Wiley-VCH, Mannheim 2000; Vol. 1, p. 667.
30. Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V.; Fresenius, J. *Anal. Chem.* **1997**, *359*, 50.
31. Akaike, H. 'Information theory and an extension of the maximum likelihood principle'; Second International Symposium on Information Theory, **1973**, Budapest.
32. Akaike, H. *IEEE Trans. Automat. Contr.* **1974**, *AC-19*, 716.
33. Kubinyi, H. *Quant. Struct. Act. Relat.* **1994**, *13*, 285.
34. Kubinyi, H. *Quant. Struct. Act. Relat.* **1994**, *13*, 393.
35. Friedman, J. In *Technical Report No. 102*. Laboratory for Computational Statistics, Stanford University: Stanford, 1990.
36. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell* **2002**, *20*, 269.